Private State in Public Media
Subjectivity in French Traditional and On-line News

Anne Küppers and **Lydia-Mai Ho-Dac**[1]

Abstract. This paper reports on ongoing work dealing with the linguistic impact of putting the news on-line. In this framework, we investigate differences in one traditional newspaper and two forms of alternative on-line media with respect to the expression of authorial stance. Our research is based on a comparable large-scale corpus of articles published on the websites of the three respective media and aims at answering the question to what extent the presence of the author varies in the different media.

1. Is it a matter of amount and mode of the author's presence?
2. Is it a matter of lexical choice and diversity?
3. If this were the case, what expressions are used in the respective media?

Our endeavour will be a methodological one. We firstly present our data, and thus describe the different news media included in our study, and the diverse computer aided and manual production steps we performed in order to build up the corpus. Secondly, we outline our working hypotheses that are linked to the chosen types of media and describe the theoretical framework within which they are situated. Thirdly, we present our research method as well as some first results and insights gained throughout the pilot study of our data.

1 Corpus

The main objective of our research is to contrast traditional newspaper language with the language used in alternative forms of journalism in order to determine whether we have to do with distinct genres, or merely different text types.

We therefore created a large-scale corpus consisting of articles published in one traditional newspaper and in two alternative written on-line mass media. Texts included in the corpus have been published between 2005 and 2009 and were collected directly from the respective website's archives. The media are briefly presented in the following and table 1 gives an overview of the sections included in each data set. Sections are chosen on the basis of comparison of the topics dealt with in order to ensure a higher degree of comparability of the different sub corpora.

The first data set consists of articles published in one of the principal Belgian, French-speaking traditional reference newspapers, namely *Le Soir*. This liberal, most read supra-regional newspaper was first published in 1887. Texts included in our corpus have been published in the printed or the on-line version of the newspaper.

The second data set is composed of articles published on the website of the French independent journalism project *Rue89*. This project started in 2007 and aims at unifying professional journalism and Internet culture. It works with a committee of professional journalists and young reporters ensuring a good portion of articles and the reviewing of texts submitted by external domain specialists.

Table 1. Newspaper Sections included in the Data Sets

Le Soir	Rue89	AgoraVox
News	World News	International News
		Europe
	Politics	Politics
	Society	Society
Culture	Culture	Culture
	Media	Media
		Religion
		Bizarre
		Life and Style
		People

The third data set is made up of articles published on the French alternative on-line information platform *AgoraVox*. This citizen press website was created in 2005 and follows the principle of editorial democracy, that is to say that any Internet user can subscribe and contribute articles. Reviews are done by members of the committee, composed of some anchormen and the sub-editors, who are potentially all members having published at least 4 articles on the AgoraVox website.

2 Data Processing

The creation of the corpus was realised through several steps. Processing procedures are different for the three media because the constitution of their on-line archives is not uniform. And, as there are no ready-made programs for automatically extracting articles from different web pages ([44]), the data collection was a first challenge. We will briefly describe the basic legs that were realised for the entire corpus.

In a first step, all articles are collected directly from the websites' archives by means of Perl scripts. The script extracts the list of articles published under a certain URL and saves them in a separate folder for each medium. A second script extracts the articles mentioned in these folders and converts the original files of different source codes into XML format. During this step, the script keeps track of information such as the medium, the section, the author, the title and the date of publication, whenever these are available. Besides, we already partially remove wire copies from our data sets, as they would falsify our analyses on the expression of authorial stance. For the same reason, interviews, poems and songs are removed man-

[1] Université catholique de Louvain (UCL), Institut Langage et Communicatiton (IL&C), Belgium, email: anne.kueppers@uclouvain.be, lydia.ho-dac@uclouvain.be

ually in a subsequent step.

All data are encoded following the TEIP5 model and are cleaned by means of computer aided and manual control sequences. The resulting files include information about the text structure and give titles, subtitles, formatted lists and paragraphs. They also indicate citations[2], bold or italic printing, a set of meta-information concerning the corpus, and each unique article. All this supplementary information is displayed by XML tags.

3 Hypotheses

On the basis of this multi-layered corpus of written mass media, we aim at bringing to light similarities and divergences in terms of linguistic and structural dimensions. Concerning the language use in the three different media, we have the following hypotheses:

1. Subjectivity is expressed differently in the three media, namely with respect to the amount and lexical choice of subjective expressions.
2. Articles in the alternative media are more subjective than those published in the traditional newspaper, as the author expresses more overtly her/his opinion and thoughts in the former ones.
3. On average, the writing style in articles published in AgoraVox is even more subjective than the style in those published in Rue89.

We formulate these hypotheses on the basis of the working processes prevailing in the editorial departments of the different media and their respective 'philosophy'. The traditional newspaper's data set is composed of two different types of articles, those published in the paper version and those published on the website. While most of the articles recorded for the printed version are based on investigation and are entirely written by one journalist, most of the articles published on the website are slightly modified wire copies that are simply adapted to the editorial line ([20]). This is a consequence of the prevailing working guideline for the web journalists: 'put the news on-line as fast as possible, be the first to publish, and get the scoop'.[3] This principle influences the writing style, which we assume to be less individual and thus less subjective. As a consequence, the presence of the on-line articles in our data set will probably have an impact on the amount of subjective expressions, their mode and lexical choice.

The two alternative media included in our research differ from the traditional newspaper in some points: First, these media are solely published on the Internet. Second, they work with professional journalists, domain specialists and non-professional Internet users. Third, they tend to cover current topics, but not to get the scoop at all costs. Fourth, the alternative media do not intend to cover all actual topics, but just those that seem relevant to the authors – either for themselves, or with respect to their audience.

Concerning their working mode and philosophy, Rue89 defines itself as being more comparable to a radio station than to a traditional newspaper with regard to reactivity, the absence of deadlines, the exchange of participants, and the informal writing style ([6]).

Comparing Rue89 and AgoraVox, the parameter of professionalism of the authors might influence the degree of subjectivity and the way it is expressed. Members of Rue89's editorial board are professional journalists and authors submitting articles to the website are professional journalists as well, or at least so called domain specialists ([6]). Topics to be dealt with in Rue89 may be suggested by Internet users, but the latter do not participate in contributing content themselves. Articles published in AgoraVox are written by professional journalists, domain specialists and non professional web users, and the review committee is composed of professional and non-professional writers.

We assume that the language style differs between the two alternative media, with Rue89 occupying an intermediate position between Le Soir and AgoraVox. More concretely, we suppose that the writing style in articles published in AgoraVox is more individual, and more subjective due to the articles written by individual web users that are not professional journalists or domain specialists. We hypothesize the intermediate position of Rue89 due to the professionalism of authors working on the project on the one hand (which is not the case for AgoraVox) and to the independent topic choice and the time and investigation for the recording of articles on the other (which is not the case for Le Soir, at least for the part of the on-line articles).[4]

Furthermore, we are interested in outlining whether we could speak of distinct genres or text types when comparing different types of news media. Accounting for the a priori difference between the three data sets, we aim to outline whether the supposed differences effectively exist. We do not intend to point out potential differences between on-line and printed journalistic genres like bulletin, reportage, editorial or comment ([1], [37]), but to detect evidence on a more general level, namely between the three types of media under consideration.[5]

As all data sets belong to the domain of journalism, we cannot presume that the three media belong to different genres. But we expect discrepancies between the data sets that are due to (1) the professionalism and education of authors writing for Le Soir, (2) the aim to diffuse objective information in the sense of reflecting news without judging in Le Soir, (3) the aim of Rue89 and AgoraVox to report differently from the traditional press, i.e. not following neither a particular editorial line, nor a given deadline for article publishing, (4) the aim to make non-professional writers participate in news coverage as is the case for AgoraVox. If the distinction of genres turns out to be too general for our purpose of distinguishing three types of journalese, we will still try to outline representative text types for each of them.[6]

Whether we can speak of different genres or text types when comparing the three media under consideration is tested by the phenomenon of speaker stance.[7] In order to outline the expression of subjectivity in our newspaper corpus, we use a twofold method including deductive and inductive quantitative approaches which are presented in section 5, as well as a qualitative analysis.

[2] As compared to interviews, single citations included in articles are kept and tagged in order to easily identify and exclude when wanted, as the choice of a citation reflects a personal state, not only of the source, but also of the author, and citations thus may serve our subsequent analyses of subjectivity.

[3] As the traditional newspaper aims to treat all actual topics also in the paper version, journalists may have time pressure due to deadlines when recording for the printed support as well, depending on the topic of the article.

[4] Texts for which this might be the case when regarding the paper version of Le Soir, such as wire copies or newsflashes, have been excluded from our corpus as explained in section 2.

[5] As studies on subjectivity are often based on corpora build up of texts dealing with the same topic, we plan to compare different sub corpora in subsequent studies.

[6] In the domain of corpus linguistics, the term *text type* was first introduced by Biber ([9]: 68) who defines it by means of inner-textual linguistic characteristics, as opposed to *register* (previously *genre*), which is defined in terms of external and cultural criteria linked to the author's purpose.

[7] To determine in a more general way whether we effectively have to do with different genres or text types will need further investigation concerning other linguistic dimensions, but our pilot study already allows for the detection of tendencies.

4 Theoretical Framework and Research Method

Our quantitative analysis is based on two axes, namely (1) discourse organisation through initial position (i.e. the first preverbal zone of a given sentence) and (2) subjectivity through PSEs. While the first focuses on the evaluation of the typological differences between the media, the second is devoted to subjectivity in order to observe the variation of the author's presence in the texts. Nevertheless, the two axes interact: Not only subjective language, but also the order of information reflects speaker stance, namely by choosing the information included, by mentioning certain aspects before others, or by linking different texts parts (phrases, sentences, paragraphs).

Before we introduce our research methods, we briefly sketch the theoretical principles our research is based on. The work in progress presented is situated in a corpus linguistic framework of discourse analysis. The present paper aims to outline first tendencies in our data. We describe and evaluate the typological differences between the three media by applying a corpus-based methodology providing a description of the global discourse organisation of our data sets and the expression of subjectivity by certain predefined cues (section 4.3).

4.1 Discourse Organisation through Initial Position

Because of the apparent incompatibility between the qualitative nature of discourse analysis and the quantitative requirements of corpus linguistics, discourse organisation is usually difficult to study by means of corpus linguistic methods ([8]). Ho-Dac [27] proposes a method providing a solution to this incompatibility, allowing for a data-driven approach to discourse organisation based on automatic tagging and quantitative analysis of the discourse roles of sentence-initial elements in different text positions given by the layout. The theoretically-based hypothesis is that the initial position – defined as the starting point of the message and composed of the first elements that the reader receives – has an important function in discourse organisation. The analysis of the distribution of these elements according to their text position gives an overview of the textual organisation of different text types. Therefore, two text positions are distinguished: *P1* corresponding to sentences introducing a paragraph, and *P2* corresponding to intraparagraphic sentences. Elements in P1 are by definition associated with a paragraph break, i.e. a visual cue of discontinuity. As a consequence, they have a greater capacity of signalling high-level discontinuities and orienting high-level segments. Because discourse organisation is complex and texts are organised according to different structuring principles, we have to consider different types of discourse segments. In this study we focus on cues that potentially signal topical continuity, rhetorical articulation, setting discontinuity, and textual discontinuity.

Topical continuity is outlined by means of co-referential grammatical subjects covering pronouns, possessive noun phrases, reiterations, and detached appositions in initial position. Several studies in cognitive linguistics showed that linguistic means available to refer to a given entity already mentioned in the text are associated with different degrees of accessibility (e.g. [41], [2], [23]). On their basis we assume that (1) co-referential expressions, especially when occurring in grammatical subject position, have an instructional meaning indicating topical continuity, and that (2) the type of this expression indicates different levels of topical continuity. For example, a first person personal pronoun in grammatical subject position indicates a strong topical continuity while reiteration may be used to reintroduce a topic or to reinforce a topical continuity when there is a discourse shift e.g. a paragraph break, a setting or a textual discontinuity cue ([46]). Another topical continuity cue is apposition, which is an attributive construction communicating supplementary information on a given sentence constituent from which it is syntactically detached. Concerning discourse organisation, and especially when occurring in initial position just before the first grammatical subject, appositions may indicate topical continuity just like to referential links ([18]), and it has been shown that the more narrative a text, the more appositions and pronouns occur in P1 ([27]).

For **rhetorical articulations** we only consider connectives occurring in absolute first position. When introducing a sentence or a paragraph, they may acquire a high-level discourse function in order to signal a rhetorical articulation taking place inside in the course of a given continuity (concerning topic or setting). Ho-Dac ([27]) shows that the more argumentative a text, the more connectives occur in initial position.[8]

Setting discontinuity is outlined by means of detached setting adverbials. When occurring in sentence-initial position, setting adverbials may orient the reader by indicating the domain of applicability within which the following proposition holds (e.g. [13], [21] and [22]). In the present study, we focus on time, space, and notional adverbials, i.e. elements which set a notion that may be a domain of knowledge (*in linguistics*), a defined object (*concerning the case of adverbials*), a specific point of view (*in line with Halliday*), etc. The text part introduced by these adverbials is labelled discourse frame and characterized by temporal, spatial, or notional homogeneity ([17]). Ho-Dac ([27]) shows that the more descriptive a text, the more setting adverbials occur in initial position.

Concerning **textual discontinuity**, we focus on sequencers (linking adverbials and grammatical subjects introducing items) that serve to indicate discourse organisation attributing limits of different text parts and information sources by explicitly indicating the position of a given segment in discourse (e.g. *Firstly,... Secondly,... Finally,... Moreover,... Besides,... etc*).

4.2 Private State

Subjectivity generally refers to the expression of personal state, covering devices of opinion, evaluation, attitude and emotion or sentiment when generally speaking. Depending on the underlying theory and the linguistic means at focus, the phenomenon is amongst others designated as *stance* ([9], [11]), *appraisal* ([38], [51]), *hedging* ([35], [29]), *commitment* ([47]), *private state* ([42]) or *evaluation* ([5]).[9] Diverse means can serve to express subjectivity in texts. Usually any subjective element is linked to its *emitter* who can either be the writer or some other person referred to or cited in the text. In the same way, subjective elements are generally linked to a *goal* that the personal state relates to. In line with Thompson and Hunston ([48]), we define private state as

> the broad cover term for the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about. That attitude may relate to certainty or obligation or desirability or any of a number of other sets of values.[10]

[8] Connectives are more often used to link several continuous clauses inside a given sentence.

[9] For further description see Bednarek ([5]).

[10] While the defined phenomenon is labelled *evaluation* by Thompson and Hunston ([48]), we use the terms *subjectivity*, *stance* and *private state* as equivalents to it in the course of this article.

Besides the lexical choice (including single words, collocations and complex phrases), also morphology and syntax can communicate a personal state in written texts.[11] It is very important to note that the discrete occurrence of a subjective expression is not by force used in its subjective meaning, which is true for objective devices as well. Depending on the context, an a priori subjective expression can be used objectively and vice versa. The distinction of subjective and objective elements thus demands more detailed and qualitative analyses. Following Wiebe et al. ([52]: 281f), we therefore speak of Potential Subjective Elements (PSE) to refer to those "linguistic element[s] that **may** be used to express subjectivity" by means of their primary meaning (our emphasis). Whether a PSE is effectively used subjectively is dependent on the context of a given utterance.

4.3 Potential Subjective Elements

By the second axis, we explore the use of Potential Subjective Elements (PSE) in the three data sets, accounting for occurrences of first person personal and possessive pronouns[12], stance adverbials, *it*-extrapositions, cleft sentences, and hapax, i.e. words that occur just once in a given data set.

(i) Ces fameuses années 68–70, qui **nous** submergent aujourd'hui, ça commence à **m'**énerver. `Rue89_2850`
(ii) **Il faut donc** pour les africains francophones abandonnés le F CFA, fabriqué en France - près de Clermont-Ferrand... `AgoraVox_2709`
(iii) Pour Alain Menand, **il est de toute façon hasardeux** de prétendre "classer" les différentes licences : [...] `Rue89_3876`
(iv) **Je** ne reviendrai pas sur les questions rhétoriques toujours aussi efficaces. `AgoraVox_3667`
(v) Alors que l'anthropologie et la sociologie ont souvent pensé les cultures selon des modèles de groupe **nous** verrons ici ce que le concept de culture doit à la prise en compte des besoins de l'individu au plan personnel [...] `AgoraVox_2913`
(vi) Vendredi dernier dans **nos** colonnes, les recteurs de l'Université libre de Bruxelles (ULB), Pierre de Maret, et de la Vrije Universiteit Brussel (VUB), Ben Van Camp, signaient une Carte blanche. `LeSoir_5359`
(vii) **On** assiste ainsi à une soirée organisée en l'honneur des Amis américains de Versailles dans la Galerie des Glaces du fameux château. `LeSoir_4341`
(viii) **C'est l'amer constat que** l'**on** peut faire soit qu'**on** y habite ou qu'**on** y arrive pour la première fois dans cette ville qui jadis, présentait fière allure. `AgoraVox_3584`
(ix) Car **il serait évidemment bien dangereux de** se replier frileusement sur les égoïsmes nationaux d'antan. `Rue89_987`
(x) **C'est d'abord parce qu'**ils gardent au début l'espoir insensé d'un miracle, et qu'ensuite il est trop tard. `AgoraVox_1451`

The use of a first person pronoun is one of the most conspicuous means used to express subjectivity as it "refers to the act of individual discourse in which it is pronounced, and by this it designates the speaker" ([7]: 226) (examples (i), (iv)-(vi)). Several linguistic investigations on speaker stance – most of them including English as (at least one of) the language(s) under investigation and focusing on academic and scientific discourse – outline the importance of the choice of personal pronouns in order to express the degree of involvement in relation to the propositional content (e.g. [24], [25], [32], [33], [34], [49]). The third person singular pronoun can fulfill the communication of subjectivity when occurring as grammatical subject, too. In French, this is the case for the third person singular pronoun *on* used as an alternative to the first person plural pronoun *nous*, and taking over the speaker-inclusive meaning (examples (vii)-(viii)).

In addition to pronouns we include stance adverbials and two special constructions that are potentially related to subjectivity: *it*-extrapositions and cleft sentences. We are especially interested in these cues of subjective language as they offer the author the possibility to express stance in an indirect way.

Stance adverbials (like *évidemment* in example (ix)) can be used for reasons linked to content (e.g. when information about a topic is not sufficiently accessible), or for interpersonal reasons (e.g. when the author does not want to impose a personal point of view to the readers) ([30], [31]). Originally, stance adverbials have been defined as a means of hedging by rendering the affiliation of an object to a certain category fuzzier ([35]), while they are accredited now a more general function, including the expression of attitude, emotion and opinion. The particularity of *it*-extrapositions is that they express a subjective meaning while at the same time communicating a certain degree of distance between the author and the propositional content (example (ix)). As Charaudeau ([14]) points out, the use of *it*-extraposition is very frequent in journalese, as these formulations seem to be less subjective, so that we speak of *constructed objectivity* ([16]: 504). Cleft sentences can also express an indirect judgement on the propositional content of a message ([36]). Their main function is to focus on an extracted element that is detached from the other sentence components in order to be emphasised (examples (viii), (x)). By choosing a cleft sentence structure, the author can communicate the accentuation of a given propositional content.

Lastly, we focus on single word occurrences, so called hapax legomena.[13] Following Wiebe et al. ([52]), we assume that one word occurrences are especially interesting when exploring subjectivity in discourse.[14] Existing studies investigating stance by a corpus-based approach are all based on English language data ([28], [15], [10], [19], [40] amongst others). Until now – to our knowledge – there are no such analyses based on French large-scale corpora.

5 Quantitative Analysis

Our quantitative analyses of the different cues presented in section 4 are based on automatic tagging which has already been used and evaluated twice ([27] and [39]).

5.1 Cues Marking

The quantitative analyses presented are based on an automatic labelling of features concerning discourse organisation on the one

[11] In spoken or direct discourse, still other indicators like intonation, gesture, and mimicry can perform this task.
[12] The occurrences include those of the third person singular pronoun *on* in the use of *nous* (*we*).
[13] Instead of focusing just on the most frequent and statistically significant word occurrences in a given data set, what has been custom in corpus linguistics for several years, more recent studies also take into account the least frequent phenomena, namely hapax ([4], [50], [3], [52]).
[14] The inclusion of hapax in large-scale corpus investigation could also be a first step to work against the criticism mentioned in qualitative research (e.g. [5]) that bottom-up data mining is not an adequate method to outline expressions of subjectivity, as it could never detect all occurrences of subjective language due to the undefined and unlimited diversity of possible formulations.

hand, and subjectivity on the other hand. This labelling is based on the results of a POS tagging (*TreeTagger*, [45]), a syntactic parser (*Syntex*, [12]), and on layout information directly extracted from the TEI encoding of the corpus (section 2).

Concerning discourse organisation, the automatic marking extracts a selection of potential organisational cues occurring in initial position, distinguishing connectives occurring in absolute first position, detached elements, and grammatical subjects. These elements are automatically characterized by their POS, their function (setting vs. textual adverbials, sequencers, appositions, etc.), the semantic meaning (e.g. temporal, spatial, and notional setting adverbials), and the properties of reiteration (when an NP's head restates a noun already mentioned in a given section). Moreover, these elements are associated with their textual position, i.e. P1 (if their host sentence introduces a paragraph) or P2 (if the host sentence is intraparagraphic).[15] The automatic characterisation is based on a Perl script (1) delimiting the first preverbal zone for all sentences, (2) identifying all syntactic blocs composing the preverbal zone (based on *Syntex* results), 3) categorising each bloc by applying a set of regular expressions associated with lexical lists concerning functional and semantic features.

Detecting the PSEs, we use the POS database in order to outline hapax legomena. Concerning cues marking, we have adapted the Perl script used for discourse organisation to extract first person pronouns (*j*e, *n*ous, and *o*n in subject position and *m*e/*m*', *m*oi, *n*ous, *s*e/*s*' in other positions), and possessive NPs (*m*on/*ma*/*mes X*, *n*os/*notre* X). Moreover, three other cues, namely stance adverbials and *it*-extrapositions as well as cleft sentences, are automatically extracted, based on the POS tagging, a lexical list for the former, and syntactic patterns for the latter.

Table 2. Extracted Cues

DISCOURSE ORGANISATION	
SETTING	Setting adverbials
SEQ	sequencers
CONNECT	connectives (coordinations, adverbs)
APPOS	appositions
COREF_r	proper nouns, definites, demonstratives, possessive and undetermined NPs with a syntactic head reiterating a noun already mentioned
COREF_p	pronoun and possessive NPs
POTENTIAL SUBJECTIVE ELEMENTS (PSE)	
STANCE	stance adverbials
LOCpro	1st person personal pronouns (including *on*)
LOCposs	NPs with 1st person possessive determiner
IT-ex	*it*-extrapositions
CLEFT	cleft sentences

5.2 Frequency Analysis

We firstly describe the main differences between the three data sets in terms of layout, discourse organisation, and occurrences of PSEs. Secondly, we expose occurrence frequencies of the diverse cues for each data set by using contingency tables (comparing the data sets two by two) and the log-likelihood ratio (henceforth LL[16]) in order to measure the significant relative frequency differences between them.

[15] For more methodological details see Ho-Dac ([27]).

[16] See [43] for details on the use of this ratio for large-scale corpus comparison.

The higher the LL value, the more significant the difference is between two frequency scores. In this study principally aiming at describing main tendencies, we only focus on LL corresponding to $p < 0.0001$ (i.e. higher than 15.13). The resulting tendencies will or will not support our hypotheses and will constitute the starting point for further detailed, quantitative and qualitative analyses.[17] Before presenting tendencies concerning subjectivity in our corpus, the next section describe the linguistic characteristics of the three media in terms of layout and discourse organisation.

6 Interim Results and Tendencies

This section presents the first results and insights gained throughout the quantitative analysis. Its main concern is to expose observed general tendencies on the basis of frequency analyses of the cues for each data set and LL statistics for their comparison.

6.1 Linguistic Characterisation of the Three Media

To give an overview of the **general characteristics** of the three data sets, we describe in the following their layout and lexical diversity. While the first has to do with discourse organisation, the second may be linked to subjectivity ([52]). Table 3 suggests different units of measurement, more or less related to layout, in order to describe their size and textual segmentation.

Table 3. Layout Segmentation

	Rue89	AgoraVox	Le Soir	total
Words	2,187,333	**3,281,208**	2,744,270	8,212,811
Headings	687	**896**	715	2,298
Articles	3,879	4,368	**5,873**	14,120
Words/Article	564	**751**	467	582
Sentences/Article	50	**68**	42	53
Paragraphs/Article	112	**14**	8	11
Sentences/Paragraph	2.7	**3.5**	2.5	2.9

AgoraVox is the largest data set concerning the overall number of words, paragraphs, and headings. It also contains the longest articles (on average 751 words/text) and longer paragraphs as compared to the two other media. In contrast, Le Soir shows shorter articles (on average 467 words/text) and paragraphs. As a consequence, Le Soir is the larger data set with respect to the total number of articles (5,873). Rue89, the smallest sub corpus, occupies an intermediate position.[18] Paragraph size may play an important role in discourse organisation, allowing for simple structures in short paragraphs as compared to longer ones.[19]

Lexical diversity is evaluated in the present by using the type/token ratio based on the idea that the more types as compared to the number of tokens, the more varied is the vocabulary. And the closer to 1 the ratio, the more lexically diverse is the data set. Lexical diversity might be linked to authorial presence and the expression of private state in the text as outlined in sections 3 and 4.3.

[17] Our research prospects include the investigation of more quantitative analyses on modal expressions, stance adverbials, and adjectives expressing subjectivity, as well as *verba dicendi et sentiendi*. We also intend to carry out qualitative analyses on occurrences of all mentioned subjectivity cues.

[18] Rue89 is the youngest media founded in 2007, explaining the comparatively smaller size.

[19] Sentence length will be analysed in subsequent investigations, being another discourse organisation factor.

Table 4. Type/Token Ratio

Rue89	AgoraVox	Le Soir
.019447	.015928	**.018259**

The type/token ratio shows that lexical diversity is more elevated in Rue89 and Le Soir, with a higher degree of diversity in Rue89.

To characterise the three media in terms of their **discourse organisation**, we first compare them concerning the frequencies of discourse organistion cues, and second with respect to the content of P1 and P2, applying the methodology described in section 4.

Table 5 gives the LL statistics for the sentences beginning with at least one organisational cue.

Table 5. Organisational Cues' Distribution – LL Statistics

Sentences with	R vs. A	A vs. S	R vs. S
All. org. cues	**116.57 (R)**	**880.27 (S)**	**262.73 (S)**
SETTING	316.12 (R)	427.01 (S)	[P2: 39.45 (S)]
SEQ	21.97 (A)[P2]	33.00 (A)[P1]	
CONNECT	16.32 (A)[P1]	38.38 (A)[P1]	
Topical cues		**647.2 (S)**	**498.37 (S)**
APPOS	72.41 (R)[P2]	1,716.75 (S)	832.41 (S)
COREF_r	91.35 (A)[P2]	[P1: 34.55 (S)]	93.73 (S)
COREF_p		[P1: 151.29 (A)]	15.64 (R)[P1]

R = Rue89, A = AgoraVox, S = Le Soir
(R,A,S) indicates corpus with overuse
[P1,P2] indicates position with overuse

LL statistics indicate diverse differences between the three media,and support our hypothesis concerning the typological difference between them. If we look at the first rows, Le Soir appears to be the media with the highest number of cues signalling discourse organisation. But this overuse is only effective for topical continuity cues and especially via appositions and reiterations as shown in the last three rows. Nevertheless, this overuse is not effective for all topical cues. Indeed, each topical cue is significantly associated with different media: appositions with Le Soir (and in a weaker proportion with Rue89), reiterations with AgoraVox and Le Soir, and pronouns and possessive NPs with Rue89. AgoraVox is the media with the lowest amount of organisational cues. Nevertheless, this weaker proportion of organisational cues must be qualified by looking at the detailed LL indicating that there are significantly more sequencers and connectives in AgoraVox. Concerning setting cues, Rue89 and Le Soir seem to be alike, being significantly more present in the first than in AgoraVox. If we now look at the columns, Le Soir emerges as the most specific data set in contrast to Rue89 and AgoraVox that are closer in terms of discourse organisation. Nevertheless, Rue89 and AgoraVox are not similar. They weakly differ for all different cues: (1) while AgoraVox prefers reiteration, Rue89 shows a higher amount of strong topical continuity devices (appositions and pronouns), (2) while AgoraVox shows more sequencers and connectives, Rue89 shows more setting adverbials, comparable to Le Soir.

Taking into account variations according to textual position (indicated by brackets), we find significantly more setting adverbials, sequencers, reiterations, and appositions in P1 and significantly more connectives, pronouns, and possessive NPs in P2 (Ho-Dac's ([27]) results are in line with our insights). It is only if we focus on variations between media in each textual position that new insights appear. Setting adverbials in Le Soir are overused only in P2 when comparing Le Soir to Rue89, i.e. setting adverbials are overused in Le Soir when they are not associated with an effective structuring power ([26]). When connectives are overused in a media, it is generally in P1. In AgoraVox vs. Rue89 and Le Soir, but also in Le Soir vs. Rue89, the difference between the three data sets concerning the use of these argumentative elements is conspicuous. Pronouns and possessive NPs are overused in Rue89 when occurring in P1, underlying global topical continuity. In contrast, it is intraparagraphic reiterations that are overused in AgoraVox.

All these observations allow for assuming that the three data sets under investigation are different. Although further analyses are needed in order to better understand the differences, we may state here that the media show more characteristics of the argumentative text type (as compared to descriptive or expository texts), considering the use of connectives in P1 (associated with argumentative text types ([27])) as compared to the use of setting adverbials in P2 (associated with descriptive text types([27])).

6.2 PSEs as lexico-syntactic elements

The present subsection describes results concerning the research axis on subjectivity by the use of potential subjective elements, outlined in section 4.3.

Table 6. PSE Distribution – Number of Sentences

Sentences with		Rue89	AgoraVox	Le Soir
STANCE	Nb	3,229	6,059	3,201
	%	*1.65*	*2.03*	*1.28*
LOCpro	Nb	29,468	42,952	24,413
	%	*15.08*	*14.38*	*9.77*
LOCposs	Nb	4,715	7,467	4,531
	%	*2.41*	*2.50*	*1.81*
IT-ex	Nb	1,219	2,814	1,103
	%	*0.62*	*0.94*	*0.44*
CLEFT	Nb	4,567	6,256	4,762
	%	*2.34*	*2.09*	*1.91*
total		195,395	298,636	249,830

Table 6 displays the overall occurrences of PSEs included in the present investigation for each data set. As can be seen, LOCpro constitutes the most prolific cue, with about 15% of sentences containing a personal pronoun refering to the first person in the alternative media and 9% in the traditional newspaper. All other cues occur much less frequently and divergences between traditional and alternative media are not that striking. *It*-extrapositions are the least frequent means for expressing subjectivity in all data sets (R = 0.62%, A = 0.94%, S = 0.44%), while first person posssessive pronouns, stance adverbials and cleft sentences occupy an intermediate position with alike frequencies in the differnt sub corpora. It is striking that the number of sentences any of the given PSE is never hiher for Le Soir than for Rue89 or AgoraVox when considering percentages.

Table 7 represents the LL realised for the selected PSEs occurring in our corpus, comparing the data sets two by two. The first striking result is that the Le Soir data set never corresponds to the one with overuse for any of the subjectivity cues under investigation. This is in line with the comparison of percentages in table 6. Second, the divergence between alternative media on the one hand and the traditional newspaper on the other is eye-catching, especially when comparing the frequency of sentences containing a first person personal

Table 7. PSE Distribution – LL Statistics

total	R vs. A	A vs. S	R vs. S
PSE		3320.99 (A)	2829.75 (R)
Sentences with	R vs. A	A vs. S	R vs. S
STANCE	90.4 (A)	460.28 (A)	103.85 (R)
LOCpro	39.19 (R)	2,395.78 (A)	2,528.77 (R)
LOCposs		297.28 (A)	188.14 (R)
IT-ex	151.86 (A)	499 (A)	69.28 (R)
CLEFT	31.47 (R)	24.2 (A)	96.66 (R)

R = Rue89, A = AgoraVox, S = Le Soir
(R,A,S) indicates the corpus with overuse

pronoun (A vs. S LL = 2,395.78 and R vs. S LL = 2,528.77). Third, the differences comparing Rue89 and AgoraVox are much less conspicuous. Rue89 displays significantly more personal pronouns (LL LOCpro = 39.19) and cleft sentences (LL = 31.47), while AgoraVox overuses stance adverbials (LL = 90.4) and *it*-extrapositions (LL = 151.86). The frequency of possessive pronouns does not differ significantly between the two alternative media, in which they are respectively overused as compared to the traditional newspaper (A vs. S LL = 297.28 and R vs. S LL = 188.14). Rue89 seems to overuse cleft sentences (R vs.A LL = 31.47 and R vs. S LL = 96.66) that generally serve to point out an element by detachment, and which may as well be an indication for a more informal language style in Rue89. In contrast, the high frequency of *it*-extrapositions in the AgoraVox data set (R vs.A LL = 151.86 and A vs. S LL = 499) reflects an overuse of these constructions commonly associated with an impersonal expression of private state. These findings may be associated with the smaller amount of first person personal pronouns in AgoraVox as compared to Rue89. The assumption of a more informal language use in Rue89 and a more impersonal expression of subjectivity linked to it ask for further investigation.[20]

6.3 PSE as Hapax

Table 8 concerning subjectivity cues analyses occurrence patterns of hapax legomena.

Table 8. Distribution and LL Statistics Concerning Hapax Legomena

Distribution		Rue89	AgoraVox	Le Soir
Token		3,131,675	5,092,708	3,836,117
Hapax	Nb	26,849	38,006	29,777
	%	*0.86*	*0.75*	*0.78*
LL statistics		R vs. A	A vs. S	R vs. S
Hapax		300.17 (R)	25.80 (S)	139.16 (R)

R = Rue89, A = AgoraVox, S = Le Soir
(R,A,S) indicates the corpus with overuse

As can be seen, their frequency is significantly higher in Rue89 than in the two other media (R vs. S LL = 139.16 and R vs. A LL = 300.17). AgoraVox shows the lowest type/token ratio (table 4: .015928) and also the lowest amount of one word occurrences as compared to the two other media (R vs. A LL = 300.17 and A vs. S LL = 25.80). But as articles published in AgoraVox have a longer mean length (table 3: A = 751.19 words/text and S = 467.27 words/text), this first tendency has to be put into perspective and controlled by further research. Because even if "people are creative when they are being opinionated" ([52]: 286), the corpus of journalese texts may show a high amount of hapax due to technical terms and specific language linked to a given subject. This might be an explanation for the low amounts of hapax and type/token ratio in AgoraVox – linked to the participation of non professional journalists publishing in this media, the highest amount of type/token ratio in Rue89, which is due to the professionalism of authors on the one hand side and the aim to report 'differently' from the traditional newspapers on the other, and the intermediate position of Le Soir, associated with a professional and thus probably more technical but less individual language use.

7 Conclusion

The present paper outlines the occurrences patterns of potential subjective elements in three different types of written mass media. In order to outline the expression of subjectivity, we carried out quantitative analyses by which we draw first tendencies to respond to our research questions and tested our hypotheses. The results show that the use of the different PSEs varies in the three data sets, and percentages (table 6) and raw frequencies (table 7) show that their use is less frequent in Le Soir than in the two alternative media, which is consistent with our first hypothesis. While articles in the alternative media seem to be alike, they clearly differ from the traditional newspaper. First tendencies support our second hypothesis as well: The strikingly higher use of first person personal pronouns in AgoraVox and Rue89 reflects an overt presence of the author in the these two media, as compared to Le Soir. The amount of the other PSEs under consideration is also slightly lower in the traditional newspaper. By contrast, our third hypothesis was not confirmed. Our data betoke that the two alternative media seem to prefer different PSEs, but we cannot declare an intermediate position for Rue89. Concerning the presence of the author in the text, it even seems to be more overt in Rue89, given the higher amount of personal pronouns and hapax. The high frequency of *it*-extrapositions in AgoraVox may indicate a subjectivity that is expressed via constructed objectivity as compared to Rue89, where overuses of cleft sentences and first person personal pronouns may be an indication for a more informal or direct expression of subjectivity. These cues will have to be investigated in subsequent research steps, including further quantitative analyses on supplementary PSE such as adjectives, *verba dicendi et sentiendi*, or modal expressions, as well as more detailed qualitative analyses with regard to the presented PSEs and the creation of different sub corpora. Furthermore, our results support our hypothesis concerning a typological difference between the three media. The results effectively indicate differences between the three data sets, being less well-defined when comparing the two alternative media, but being conspicuous when opposing the former to the traditional newspaper.

ACKNOWLEDGEMENTS

This research was supported by grant number ARC 08-12-001 from the Belgian French Speaking Community. We thank Liesbeth Degand and Anne Catherine Simon for proofreading and useful re-

[20] As the occurrences of the different PSEs do not vary conspicuously, neither concerning their amount within the three data sets, nor when comparing the sub corpora concerning a given cue, we intend to carry out qualitative analyses for all of them. We expect from this detailed investigation insights concerning the mode of subjectivity expression (formal vs. informal) and the judgement's value (positive, negative, neutral) in order outline distinguishing means for the three media.

marks as well as the three anonymous reviewers for their critical but notwithstanding constructive comments.

REFERENCES

[1] J.-M. Adam, 'Genres de la presse écrite et analyse de discours', *Semen*, **13**, 9–15, (2001).

[2] Mira Ariel, *Accessing Noun Phrase Antecedents*, London: Routledge, 1990.

[3] Harald Baayen, *Word Frequency Distributions*, Dordrecht: Kluwer Academic, 2001.

[4] Harald Baayen and Richard Sproat, 'Estimating lexical priors for low-frequency morphologically ambiguous forms', *Computational Linguistics*, **22**(2), 155–166, (1996).

[5] Monika Bednarek, *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*, Continuum, 2006.

[6] Francoise Benhamou, Julie Lambert, and Marc-Olivier Padis, 'Le journalisme en ligne: Transposition ou réinvention? entretien avec laurent mauriac et pascal riché', *Esprit*, **3-4**, (2009).

[7] Emile Benveniste, *Problems in General Linguistics*, Coral Gables: University of Miami Press, 1971.

[8] D. Biber, U. Connor, and T.A. Upton, *Discourse on the move, using corpus analysis to describe discourse structure*, volume 28 of *Studies in corpus Linguistics*, John Benjamins Publishing Company: Amsterdam/Philadelphia, 2007.

[9] Douglas Biber, *Variation Across Speech and Writing*, Cambridge: Cambridge University Press, 1988.

[10] Douglas Biber and Edward Finegan, 'Styles of stance in english: Lexical and grammatical marking of evidentiality and affect', *Text*, **9**, 93–124, (1989).

[11] Douglas Biber, Stig Johansson, Geoffry Leech, Susan Conrad, and Edward Finegan, *Longman Grammar of Spoken and Written English*, London: Longman, 1999.

[12] Didier Bourigault, *Un Analyseur Syntaxique Opérationnel : SYNTEX*, Ph.D. dissertation, Mémoire d'HDR en Sciences du Langage, CLLE-ERSS, Toulouse, France, 2007.

[13] Wallace Chafe, *Subject and Topic*, chapter Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View, 25–55, New York/San Francisco/London: Academic Press, 1976.

[14] Patrick Charaudeau, 'Discours journalistique et positionnements énonciatifs. frontières et dérives', *Semen*, **22**, (2006).

[15] Maggie Charles, '"this mystery...": A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines', *Journal of English for Academic Purposes*, **2**, 313–326, (2003).

[16] Maggie Charles, 'Construction of stance in reporting clauses: A cross-disciplinary study of theses', *Applied Linguistics*, **27**(3), 492–518, (2006).

[17] Michel Charolles, 'L'encadrement du discours; univers champs domaines et espaces', *Cahier de Recherche Linguistique, LanDisCo université Nancy2*, (6), (1997).

[18] Bernard Combettes, 'Les constructions détachées comme cadres de discours', *Langue Francaise*, **148**, 31–44, (2005).

[19] Susan Conrad and Douglas Biber, *Evaluation in Text. Authorial Stance and the Construction of Discourse*, chapter Adverbial Marking of Stance in Speech and Writing, 56–73, Oxford: Oxford University Press, 2000.

[20] Amandine Degand, 'Le multimédia face à l'immédiat', *Communication*, (submitted).

[21] Simon C. Dik, *Theory of Funtional Grammar Complex and Derived Constructions*, Berlin/New York: Mouton de Gruyter, 1997.

[22] Peter Fries, *On Subject and Theme: A Discourse Functional Perspective*, chapter Themes Method of Development and texts, 317–359, John Benjamins: Amsterdam/Philadelphia, 1995.

[23] Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski, 'Cognitive status and the form of referring expressions in discourse', *Language*, **69**, 274–307, (1993).

[24] Nigel Harwood, ''nowhere has anyone attempted...in this article i aim to do just that': A corpus-based study of self-promotional i and we in academic writing across four disciplines', *Journal of Pragmatics*, **37**, 1207–1231, (2005a).

[25] Nigel Harwood, ''we do not seem to have a theory...the theory i present here attempts to fill this gap': Inclusive and exclusive pronouns in academic writing', *Applied Linguistics*, **26**(3), 343–375, (2005b).

[26] L.-M. Ho-Dac and M.-P. Pry-Woodley, 'Mthodologie exploratoire outille pour l'tude de l'organisation du discours', in *Actes du Congrs Mondial de Linguistique Franaise (CMLF-08)*, Paris, (2008).

[27] Lydia-Mai Ho-Dac, *A Mosaic of Corpus Linguistics. Selected Approaches.*, chapter An exploratory data-driven analysis for describing discourse organization, 79–100, Frankfurt/Berlin: Peter Lang, 2010.

[28] Susan Hunston and Geoff Thompson, *Evaluation in Text. Authorial Stance and the Construction of Discourse*, Oxford: Oxford University Press, 2000.

[29] Ken Hyland, 'Hedging in academic writing and eap textbooks', *English for Specific Purposes*, **13**, 239–256, (1994).

[30] Ken Hyland, 'Writing without conviction? hedging in science research articles', *Applied Linguistics*, **17**, 433–454, (1996).

[31] Ken Hyland, 'Persuasion and context: The pragmatics of academic metadiscourse', *Journal of Pragmatics*, **30**, 437–455, (1998).

[32] Ken Hyland, 'Stance and engagement: A modal of interaction in academic discourse', *Discourse Studies*, **7**(2), 173–192, (2005).

[33] Ken Hyland and Polly Tse, 'Metadiscourse in academic writing: A reappraisal', *Applied Linguistics*, **25**(2), 156–177, (2004b).

[34] Chih-Hua Kuo, 'The use of personal pronouns: Role relationships in scientific journal articles', *English for Specific Purposes*, **18**(2), 121–138, (1999).

[35] George Lakoff, 'Hedges: A study in meaning criteria and the logic of fuzzy concepts', in *Papers from the Eighth Regional Meeting*, eds., Paul Peranteau, Judith Levi, and Gloria Phares, pp. 183–228. Chicago Linguistics Society (CLS 8), (1972).

[36] Knud Lambrecht, *Language Typology and Language Universals*, chapter Dislocation, 1050–1079, Berlin/New York: Mouton de Gruyter, 2001.

[37] Gilles Lugrin, 'Le mélange des genres das l'hyperstructure', *Semen*, **13**, (2001).

[38] James R. Martin, 'Reading positions/positioning readers: Judgement in english', *Prospect: a Journal of Australian TESOL*, **10**, 27–37, (1995).

[39] M.-P. Péry-Woodley, N. Asher, P. Enjalbert, F. Benamara, M. Bras, C. Fabre, S. Ferrari, L.-M. Ho-Dac, A. Le Draoulec, Y. Mathet, P. Muller, L. Prévot, J. Rebeyrolle, L. Tanguy, M. Vergez-Couret, L. Vieu, and A. Widlöcher, 'Annodis : une approche outillée de l'annotation de structures discursives', in *TALN 2009*, Senlis, (June 2009). ATALA, LIPN.

[40] K Precht, 'Stance moods in spoken english: Evidentiality and affect in british and american conversation', *Text*, **23**, 239–257, (2003).

[41] Ellen Prince, *Radical Pragmatics*, chapter Toward a Taxonomy of Given-New Information, 223–255, New York: New York Academic Press, 1981.

[42] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik, *A Comprehensive Grammar of the English Language*, London: Longman, 1985.

[43] Paul Rayson and Roger Garside, *Comparing corpora using frequency profiling*, 1–6, October 2000.

[44] Marina Santini, 'Web pages, text types, and linguistic features: Some issues', *ICAME Journal*, **4**, 67–86, (2006).

[45] Helmut Schmid, *TreeTagger*, IMS, Universität Stuttgart, Germany.

[46] Catherine Schnedecker, *Noms Propres et Chaînes de référence*, Université de Metz : Metz, 1997.

[47] Michael Stubbs, 'A matter of prolonged fieldwork: Notes towards a modal grammar of english', *Applied Linguistics*, **7**, 1–25, (1986).

[48] Geoff Thompson and Susan Hunston, *Evaluation in Text. Authorial Stance and Construction of Discourse*, chapter Evaluation: An Introduction, 1–27, Oxford: Oxford University Press, 2000.

[49] Irena Vassileva, 'Who am i/who are we in academic writing?', *International Journal of Applied Linguistics*, **8**(2), 163–190, (1998).

[50] Marc Weeber, Rein Vos, and R. Harald Baayen, 'Extracting the lowest-frequency words: Pitfalls and possibilities', *Computational Linguistics*, **26**(3), 301–318, (2000).

[51] Peter R. R. White, *Appraisal Outline*, www.grammatics.com/appraisal, 2001.

[52] Janyce Wiebe et al., 'Learning subjective language', *Computational Linguistics*, **30**(3), 277–308, (2004).